Published in final edited form as:

Contemp Clin Trials. 2013 March ; 34(2): 248–256. doi:10.1016/j.cct.2012.11.006.

On the use of robust estimators for standard errors in the presence of clustering when clustering membership is misspecified

Manisha Desai^{a,*}, Susan W. Bryson^b, and Thomas Robinson^c

^aDepartment of Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

^bDepartment of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA, United States

^cDivision of General Pediatrics, Department of Pediatrics and Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

Abstract

This paper examines the implications of using robust estimators (REs) of standard errors in the presence of clustering when cluster membership is unclear as may commonly occur in clustered randomized trials. For example, in such trials, cluster membership may not be recorded for one or more treatment arms and/or cluster membership may be dynamic. When clusters are well defined, REs have properties that are robust to misspecification of the correlation structure. To examine whether results were sensitive to assumptions about the clustering membership, we conducted simulation studies for a two-arm clinical trial, where the number of clusters, the intracluster correlation (ICC), and the sample size varied. REs of standard errors that incorrectly assumed clustering of data that were truly independent yielded type I error rates of up to 40%. Partial and complete misspecification of membership (where some and no knowledge of true membership were incorporated into assumptions) for data generated from a large number of clusters (50) with a moderate ICC (0.20) yielded type I error rates that ranged from 7.2% to 9.1% and 10.5% to 45.6%, respectively; incorrectly assuming independence gave a type I error rate of 10.5%. REs of standard errors can be useful when the ICC and knowledge of cluster membership are high. When the ICC is weak, a number of factors must be considered. Our findings suggest guidelines for making sensible analytic choices in the presence of clustering.

Keywords

Robust variance estimators; Sandwich estimators; Clustered data; Correlated errors; Randomized clinical trial

1. Introduction

Clustered randomized trials are studies where measurements are taken from subjects or items within groups. This may be a consequence of the sampling scheme and/or the

treatment allocation. An example is a group randomized clinical trial evaluating the effect of an educational program on the prevention of drug abuse, where the intervention is administered to schools randomized to receive either the novel program or standard education. Students from the same school may be influenced more similarly than students from different schools, particularly if the educational teams vary by school.

In many trials, cluster membership may be clear, as in the trial described above, where children attending the same school are considered to belong to the same cluster. Defining clusters in some trials, however, may be more complicated. Membership may be dynamic (for example, if in the trial above students were to change schools during the study), or students may belong to more than one cluster. Consider Stanford GOALS, a member of the Childhood Obesity Prevention and Treatment Research (COPTR) consortium, a clinical trial where the intervention at one study center is multi-faceted and includes subject participation in after school team sports as well as multiple visits by staff members to modify the home environment, among other intervention components. In this case, subjects may be clustered through interactions with other subjects during after school sports as well as through sharing of the staff members who conduct home visits, and therefore, at any given time point, a subject could belong to multiple clusters. Additionally, the clustering structure could vary over the study period through changes in both after school groups and staff members. Finally, cluster membership may be inaccessible for a proportion of subjects. For example, in Stanford GOALS, while controls subjects are encouraged to participate in health education activities at community centers, their attendance is not recorded, making it challenging to define how control subjects may be clustered.

Assuming independence across subjects in the presence of clustering can lead to biased estimates of standard errors [1–4]. Typically if the intraclass correlation (ICC) is positive (items from the same cluster behave more similarly than items from different clusters) and the explanatory measure varies within cluster (e.g., weight measurements over time from the same individual), the standard errors estimated assuming independence will overestimate the true standard error, whereas if the explanatory variable is constant within cluster (e.g., treatment assignment of students within a school randomized to study arm) the standard errors will be underestimated. Incorrect standard error estimates can result in invalid test statistics and confidence intervals, thereby impacting inference [1–4]. Underestimated standard errors could cause the null hypotheses to be falsely rejected leading one to incorrectly conclude that an association exists (inflated type I error), whereas overestimated standard errors could lead to an inability to observe an interesting effect (increased type II error).

When designing group-randomized trials, Murray and colleagues urge consideration of both the ICC and the cluster size (CS) [4]. More specifically, the increase in between-cluster variance due to the ICC can be estimated by a quantity referred to as the design effect (DEFF). This is a function of both the ICC and the CS. Expressed as $1 + (CS-1)ICC$, a desirable DEFF is one that is as close to 1 as possible. A small ICC alone does not translate to a small DEFF if the CS is large. Large DEFFs indicate a need to increase the sample size or decrease the CS, if possible, in order to attain desired precision.

Robust estimation is a tool that is commonly applied to appropriately estimate standard errors in the presence of clustering. Also known as sandwich estimators, robust estimators were developed by Huber [5] and introduced to the econometrics world by White [5,6]. Robust estimation of standard errors is a general method of empirically correcting variance-covariance estimates in the presence of clustering. Specifically, suppose β represents the vector of parameters of interest from a regression framework. Then let

$$\widehat{V} = - \left(\frac{\partial^2 \ln L}{\partial \widehat{\beta} \partial \widehat{\beta}} \right)^{-1}$$

represent the information matrix that, under the usual regularity conditions, is consistent for V , the true covariance matrix. Then the robust estimator used to empirically correct for clustering can be expressed as

$$\widehat{V}_R = \widehat{V} \sum_{i=1}^m \left[\left(\sum_{j=1}^{n_i} \widehat{u}_{ij} \right) \left(\sum_{j=1}^{n_i} \widehat{u}_{ij}^2 \right) \right] \widehat{V}$$

where m is the number of clusters, n_i is the number of items in the i th cluster and $\widehat{u}_{ij}$ is the contribution to the scores, $\frac{\partial \ln L}{\partial \widehat{\beta}}$, evaluated at $\widehat{\beta}$ [1,5,6]. When cluster membership is known and the number of clusters is sufficiently large, such estimators have desirable properties [7,8]. Performance of REs when cluster membership is misspecified has not been characterized, however. In our study we 1) examine whether results are sensitive to misspecification of the clusters and 2) characterize the performance of robust standard error estimates by level of misspecification and strength of ICC and/or DEFF. We specifically focus on situations where the treatment effect is of interest, and clustering is nested within treatment arm in a randomized clinical trial. In such situations, one would anticipate underestimation of the standard errors if the clustering were ignored.

2. Design of Simulation Study

Let m be the number of clusters, n the number of subjects per arm, and N the total sample size. We conducted 31 simulation studies for a two-arm study, where the number of clusters varied from no clusters (denoted as $m=0$) to $m=4$, 20, and 50 clusters, where the sample size per group varied ($n=50$, 100, 500), where a treatment effect was and was not present ($\beta=0$ or $\beta=0.5$), where the ICC varied ($\text{ICC}=0$, 0.001, 0.01, 0.09, 0.20, and 0.50), and where the DEFF varied (1–25.5). One thousand simulated data sets were generated for each scenario. Table 1 provides a description of the scenarios. Unclustered data (Scenarios A–F) were generated from independently and identically normally distributed errors with mean zero and a variance of 1. Clustered data (Scenarios G–EE) were generated from a linear mixed effects model where the cluster-specific random intercept and random errors were each generated from normal distributions, and where clustering was nested within study arm.

Table 2 lists the 31 different clustering memberships assumed. The first 25 clustering structures (Clusters 1–25) were constructed with no consideration given to the true cluster membership and are therefore referred to as completely misspecified clusters. Clusters 26–31 were constructed to retain the original membership with probabilities that ranged from 0.70 (Cluster 31) to 0.95 (Cluster 26). The latter are referred to as partially misspecified clusters. For each data set, in addition to the true model (linear regression model for unclustered data and mixed effects model for clustered data), a linear regression using ordinary least squares methods (OLS) was fitted and standard errors were robustly estimated under each of the first 25 clustering assumptions listed. Additionally, if the data were generated from clusters, an OLS model was fitted assuming independence and with standard errors robustly estimated under the true cluster structure and under 6 clustering structures (Clusters 26–31) that partially misspecify cluster membership (Scenarios G–EE).

For data generated under no treatment effect (Scenarios A–C and Scenarios G–EE), performance of the robust estimators was gauged by the percentage of time the null hypothesis was rejected by applying a two-sided Wald test that there is no treatment effect at the 0.05 level of significance. For data generated under a treatment effect (Scenarios D–F), the percentage of time the 95% confidence interval contained the true parameter value was computed to evaluate performance. Thus, for the former, the nominal type I error was 5%, and for the latter, the nominal coverage percentage was 95%.

3. Results

In studies where participants may be clustered by design features (e.g., by administration of intervention) or by sampling scheme (e.g., by families) one does not always know whether the clustering generates data with non-zero intracluster correlations ($ICC = 0$). First we assess the effects of assuming correlation within clusters when it does not exist, with and without a treatment effect.

3.1. Misspecification of clusters for unclustered data with no treatment effect

Table 3 presents results of fitting models to data generated under Scenario A where there were 50 subjects per treatment arm ($N=100$), no clustering ($m=0$), and no treatment effect. The true model falsely rejected the null hypothesis in 5.1% of simulations, as expected. Cluster 1 assumes no clustering (i.e., that each subject is its own cluster) and yielded a similar type I error rate of 5.4%. The type I error rate corresponding to the other assumptions varied from 5.9% to 38.9%. In general, performance improved as the number of assumed clusters increased. This is to be expected as increasing the number of clusters approaches the true underlying clustering membership (no clustering = each subject as its own cluster). Even within a given number of assumed clusters, however, performance varied where variability was higher for fewer numbers of clusters and depended on features such as whether the clustering was specified as nested across arms and whether the assumed cluster size was approximately balanced (Fig. 1). When 4 clusters were assumed, the type I error rate ranged from 18.2 to 38.9%. Even when a sufficiently large number of clusters was assumed ($m=50$ or greater), the type I error rates varied from 5.9% to 11.1% and the performance was worse when clusters were assumed to be allocated differentially by treatment arm. Clustering that assumed nested and unequal membership performed the worst. The patterns observed in Scenario A persisted with increased sample size as in Scenarios B and C (shown in Table A1 in the appendix).

3.2. Misspecification of clusters for unclustered data with treatment effect

Table 3 also shows results from robustly estimating standard errors when the data are unclustered, there is a treatment effect ($\beta=0.5$) and the sample size is 50 subjects per arm (Scenario D). Because the null hypothesis is not true, we gauge performance using the percentage of time the true effect is contained in the 95% confidence interval. This is 94.8% and 94.7% for the true model and the model that assumed each item was its own cluster (Cluster 1), respectively. Models that assumed only 4 clusters covered the true parameter only 62.1%–81.4% of the time. Assuming 10 clusters yielded coverage of the true parameter 84.3%–90.2% of the time, and assuming 50 clusters or greater improved performance where coverage ranged from 89.2%–94.3%. As in the case where there was no treatment effect, models assuming clusters closest to the true structure performed the best. Increasing the sample size did not alter these results (see Scenarios E and F in Table A1 in the appendix).

Next, we examine the effects of using robust standard errors when data are generated from clusters with non-zero intracluster correlations ($ICC = 0$) but the cluster membership is misspecified. Specifically we investigate the impact of misspecifying cluster membership on

the performance of robust standard errors by 1) the number of true clusters and 2) the strength of the ICC.

3.3. Misspecification of membership by number of clusters for strongly clustered data (ICC=0.50)

Table 4 presents results from simulations where data were generated with $N=100$, with a strong ICC (0.5) and with a varying number of clusters ($m=4, 20$, and 50). Although the use of robust standard error estimates is only recommended when there are a large number of clusters (>40) [1], we included in our assessment scenarios where data arise from a smaller number of clusters. For all values of m , fitting the true model from which the data were generated (a mixed effects model) yielded desirable type I error rates of 4.4%, 5.1% and 4.6%, respectively. As expected when the data arose from a small number of clusters ($m=4$), however, performance of the robust standard error estimates was poor (Scenario G) even when the true clustering structure was specified, which yielded an inflated type I error rate of 31.6% (See Fig. 2). This markedly improved as the number of clusters increased (7.4% and 7% for $m=20$ and 50 , respectively). Similarly, ignoring the correlation had greater implications for smaller numbers of clusters, with a type I error rate of 63% for $m=4$, 29.3% for $m=20$, and 17.3% for $m=50$. In addition, misspecification of clusters had worse consequences for smaller m . Completely misspecifying cluster membership (Clusters 1–25) for $m=4$ gave corresponding type I error rates that ranged from 63.2% to 81.6% while partial misspecification (Clusters 26–31) yielded type I errors that ranged from 33.9%–46.5%. In contrast, for $m=20$, completely misspecified clusters had type I error rates that varied from 29.9% to over 60%. Partial misspecification yielded inflated but markedly lower type I errors that ranged from 8.4%–16.4%. Finally, when the true number of clusters was sufficiently large ($m=50$), performance of models using robust standard errors was improved (Scenario I). Completely misspecified membership, however, yielded a range of inflated type I error rates of 17.8%–51.7%. Assuming partially misspecified clusters provided type I error rates that ranged from 7.6%–11.6%.

For all values of m , a linear regression model that assumed independence performed better than those that made use of robust standard error estimates when information regarding the clusters was completely ignored (Clusters 1–25), while partial misspecification (Clusters 26–31) led to a reduction in the inflation of the type I error rates. Results were comparable when the sample size was increased (see Table A2 in the appendix).

3.4. Misspecification of membership by strength of ICC for a large number of clusters ($m=50$)

Table 5 presents results from data generated from a large number of clusters ($m=50$) for moderate to weak ICCs (0.20, 0.09, 0.01, and 0.001). As expected, fitting the true model from which the data were generated yielded nominal type I error rates under all four conditions. An OLS model that ignored the clustering produced corresponding type I error rates of 10.5%, 7.3%, 4.8% and 5.7% for ICC=0.20, 0.09, 0.01, and 0.001, respectively. Interestingly, robust standard error estimates with the correct cluster membership performed similarly (6.8%, 6.8%, 5.2% and 7.5%). Completely misspecified clusters yielded type I error rates that ranged from 10.5% to 45.6% for ICC=0.20, 7.3% to 39.4% for ICC=0.09, 4.9% to 39.1% for ICC=0.01, and 6% to 37.6% for ICC=0.001. Partially misspecified clusters fared better with ranges of 7.2% to 9.1% for ICC=0.20, 6.8% to 7.8% for ICC=0.09, 5.2% to 6.4% for ICC=0.01, and 6.9% to 7.8% for ICC=0.001. Fig. 3 graphically depicts the difference in distance between the type I error rate when clustering is ignored and when the true model is fitted across moderate to low ICCs. When the ICC=0.20 (Scenario O) there was a large difference in type I error between the true model and that which ignores clustering. Use of robust estimators that partially misspecify membership improved

considerably upon ignoring the clustering. When the ICC=0.09 (Scenario U), however, ignoring the clustering produced a type I error that was comparable to those obtained from partially misspecifying membership and only slightly higher than that from the true model. When the ICC was low (ICC=0.01 and ICC=0.001; Scenarios AA and EE), the benefit of incorporating partial knowledge into the clustering membership was no longer apparent; distance between the type I error rate from ignoring clustering and that from the true model was even further reduced, whereas the distance between the type I error rate from the partially misspecified and true models was similar to that obtained with more moderate ICCs (ICC=0.20 and ICC=0.09).

3.5. Level of misspecification of clustered data by distance from true number of clusters and by design effect

Fig. 4 shows the relationship between the distance between the true and assumed number of clusters and the type I error rate for Scenario I ($m=50$, ICC=0.50, $N=100$). Even when the true number of clusters was specified, the type I error rate ranged from 8% to 22%. To further evaluate the effect of level of misspecification we evaluated the variability in the type I error rate as a function of the partially misspecified clusters, each of which assumed the correct number of clusters, but varied in probability of correctly classifying items. Cluster 26 classified each item in its true cluster with a high probability (0.95), whereas Cluster 31 classified each item in its true cluster with a lower probability of 0.70. As membership assumptions improved, the type I error rate converged to the nominal rate. For completely misspecified clusters, a subtle negative relationship was observed between the distance from the true number of clusters assumed and the type I error rate, where assuming more clusters than true resulted in a lower false positive rate than the reverse.

Fig. 5a and b describes the relationship between the type I error rate and the DEFF, a function of both the ICC and the average size of the cluster. Fig. 5a shows this relationship across models for all values of the DEFF and Fig. 5b zooms in on the relationship for DEFFs less than 2. A label on the bottom of the figures indicates those observations that correspond to data generated from 4 clusters; nonlabeled observations correspond to data generated from 20 or more clusters. As expected fitting the true model from which the data arose yielded nominal type I error rates. Fitting a model that incorrectly ignored correlation had a linearly increasing impact on the type I error as the DEFF increased. For models that completely misspecified classification, this relationship held, and furthermore, these models performed worse than assuming independence for all DEFFs. For models that partially misspecified classification, the story was more nuanced. For DEFFs < 1.5 (Fig. 5b), a model that assumed independence performed comparably to one that partially misspecified classification, and in cases where the data arose from a small number of clusters ($m=4$), it was superior (in Fig. 5b, these points correspond to the spiked peaks at DEFF=1.02 and 1.24 indicated by the $m=4$ label on the plot). For DEFFs that are 1.5 or greater, the benefit of partial knowledge over assuming independence was apparent.

4. Discussion

The consequences of ignoring a clustered study design in the analysis have been well established [1–3]. Robustly estimating standard errors [5,6] is one among several approaches including mixed effects and generalized estimating equations techniques that can be used to address clustering [1,7,8]. A search in PubMed for articles with key words of “robust standard error”, “robust variance”, or “sandwich estimator” demonstrated a marked increase in their use over time. Such articles increased from 8 in the period spanning 1997–1999 to about 30 in 2003–2005 to over 100 in 2009–2011. One reason for their rise in use may be that the tools to incorporate such estimators are increasingly more accessible and incorporated into most standard software packages, particularly for use in the presence of

clustering. Another reason for their gain in popularity may be the robustness property, which could lead to a false sense of security that even if the specification of the model is incorrect, regardless of the conditions, inference is valid. David Freedman warns of the bias that can result by drawing inference from a parameter that may not be relevant under a misspecified model [9].

While the properties of robust estimators have been well investigated in various settings by others [7–14], the behavior of these estimators has not been studied when clusters are not well defined. For many clinical trials, the clustering of items is clear. For example, longitudinal studies of individuals over time induce correlation of errors from measurements arising from the same individuals. Educational interventions delivered to entire schools would likely incur correlation of measurements from students within the same school, and studies that include data on family members would likely expect responses to be clustered by family. There are studies, however, where cluster membership is less clear. The Stanford GOALS trial described previously presents such a challenge that includes membership to multiple clusters, potential membership changes over time, and inability to track cluster assignment for the control subjects. A study by Zorn [14] discovered that how the unit of clustering was defined resulted in varied estimates of standard errors when they were robustly estimated. While Zorn did not investigate misspecification of cluster membership or the role of the ICC, he found that assumptions about the unit on which the data were clustered had greater impact on inference than decisions about types of estimators.

The implications of misspecifying cluster membership in our study, which focused on scenarios where clustering was nested within treatment arm in a clinical trial setting, are a high probability of falsely rejecting the null hypothesis. The intuition behind the inflated type I error rates is as follows. Recall that if the ICC is positive and the explanatory variable (in our case, treatment arm assignment) is constant within cluster, the naïvely estimated standard error that ignores the clustering structure would underestimate the true standard error. The robustly estimated standard error exploits the empirical covariance structure based on the assumed cluster membership to improve upon the naïve standard error estimate to therefore yield an estimate that is larger than that of the naïve. If, however, there is less heterogeneity across clusters under an incorrectly assumed membership structure than under the true structure, the robustly estimated standard error will be smaller than the naïvely estimated standard error (and importantly, smaller than the true standard error), resulting in inflated type I error rates.

Our study points to the need to make careful and thoughtful assumptions about both the correlation and clustering structure in choosing an analysis strategy for potentially clustered data. Given a sufficient number of clusters, robust estimators of standard errors can be useful when knowledge of cluster membership is high (whether the ICC is strong or weak). When the ICC is non-existent such that the errors corresponding to subjects' outcomes are uncorrelated, results can be misleading by incorrectly assuming correlation exists. Similarly, even if subjects are clustered and the number of clusters is sufficiently large, results can be misleading if clustering membership is misspecified. In these cases, an OLS model that assumes independence is superior to one that completely misspecifies the structure.

In many cases, however, one is likely to have some information on cluster membership. Consider a randomized trial where an educational program to increase physical activity is delivered in a group format while the control group is encouraged to utilize a community center on their own. Subjects may therefore be considered clustered by physical activity group if assigned to the intervention or by community center if assigned to the control. If attendance of the physical activity groups is recorded and attendance at the community center is not, only partial information on cluster membership may be available. Partial

misspecification was found to produce lower type I error rates than assuming independence when the ICC is strong and given a sufficient number of clusters. When the data are weakly clustered in an unclear manner, however, assuming independence produces comparable (and in some cases, superior) type I error rates to partial misspecification, provided that the DEFF is not 1.5 or larger. Thus, when the ICC is weak and confidence in membership is not high, we recommend assuming independence. Finally, if data are generated from a small number of groups, robust estimation of the standard errors even under correct specification is not recommended. In this case, it is even more important that one try to determine the true conditions from which the data were generated in order to explicitly model the group structure, for example using techniques such as mixed effects models.

Our study has demonstrated that misspecification of clustering membership has important implications. While robust standard error estimates are known to produce statistically valid results when the correlation structure is misspecified and the number of clusters is sufficiently large [1,4], this holds only when the clustering membership is known. Our study demonstrates that misspecification of the clustering membership matters, even when the data arise from a sufficiently large number of clusters. Specifically robust standard errors can be used to account for clustering present in the data and are robust to misspecification of the correlation structure of the clusters. They are not, however, robust to misspecifications of cluster membership.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Professor David Murray for his thoughtful comments and insight on our work. The work described was supported in part by the Child Health Research Institute at Stanford University and Award Number U01HL103629 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The Sponsors had no role in study design, data collection, analysis, interpretation of data, writing of the report, or the decision to submit the paper for publication.

References

1. Diggle, P.; Liang, K-Y.; Zeger, SL. Analysis of longitudinal data. New York, NY: Oxford University Press; 1994.
2. Vonesh, EF.; Chinchilli, VM. Linear and nonlinear models for the analysis of repeated measurements. New York, NY: Marcel Dekker; 1997.
3. Donner, A.; Klar, N. Design and analysis of cluster randomization trials. London, England: Arnold; 2000.
4. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004; 94(3):423–432. [PubMed: 14998806]
5. Huber, PJ. The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; University of California Press; Berkeley, CA. 1967. p. 221-233.
6. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980; 48(4):817–838.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.
8. Liang KY, Zeger SL, Qaqish B. Multivariate regression analysis for categorical data. J R Stat Soc Ser B. 1992; 54(1):3–40.

9. Freedman DA. On the so-called “huber sandwich estimator” and “robust standard errors”. *Am Stat.* 2006; 60(4):299–302.
10. Carroll RJ, Wang S, Simpson DG, Stromber AJ, Ruppert D. The sandwich (robust covariance matrix) estimator. Technical report. 1998 <http://www.stat.tamu.edu/~carroll/ftp/sandwich.ps>.
11. Wu CFJ. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat.* 1986; 14(4):1261–1295.
12. Breslow N. Test of hypotheses in overdispersion regression and other quasiliikelihood models. *J Am Stat Assoc.* 1990; 85(410):565–571.
13. Rivals I, Personnaz L. No free lunch with the sandwich [sandwich estimator]. *IEEE Trans Neural Netw.* 2003; 14(6):1553–1559. [PubMed: 18244599]
14. Zorn C. Comparing GEE and “Robust” standard errors for conditionally dependent data. *Polit Res Q.* 2006; 59(3):329–341.

Fig. 1.
Type I error rate for scenario A. ICC=0.0; No Clusters ($m=0$); $N=100$.

Fig. 2.

Type I error rate for scenario G: ICC=0.50; 4 Clusters ($m=4$); $N=100$.

Fig. 3.
Type I error rate assuming. True, independent, and partially misspecified models by ICC.

Fig. 4. Type I error rate by distance from true number of clusters for scenario I: ICC=0.50; 50 clusters ($m=50$); $N=100$.

Fig. 5.

a. Type I error rate by design effect under various models. b. Type I error rate by design effect under various models for design effects less than 2.

Table 1

Description of simulation studies.

Scenario	Intraclass correlation (ICC)	Treatment effect	Model from which data were generated	Number of clusters (<i>m</i>)	Sample size per group (<i>n</i>)	Design effect 1+ICC(CS-1)
A	0	None (beta=0)	Linear	0	50	1.00
B	0	None (beta=0)	Linear	0	100	1.00
C	0	None (beta=0)	Linear	0	500	1.00
D	0	Yes (beta=0.5)	Linear	0	50	1.00
E	0	Yes (beta=0.5)	Linear	0	100	1.00
F	0	Yes (beta=0.5)	Linear	0	500	1.00
G	0.5	None (beta=0)	Mixed effects model	4	50	13.00
H	0.5	None (beta=0)	Mixed effects model	20	50	3.00
I	0.5	None (beta=0)	Mixed effects model	50	50	1.50
J	0.5	None (beta=0)	Mixed effects model	4	100	25.50
K	0.5	None (beta=0)	Mixed effects model	20	100	5.50
L	0.5	None (beta=0)	Mixed effects model	50	100	2.50
M	0.2	None (beta=0)	Mixed effects model	4	50	5.80
N	0.2	None (beta=0)	Mixed effects model	20	50	1.80
O	0.2	None (beta=0)	Mixed effects model	50	50	1.20
P	0.2	None (beta=0)	Mixed effects model	4	100	10.80
Q	0.2	None (beta=0)	Mixed effects model	20	100	2.80
R	0.2	None (beta=0)	Mixed effects model	50	100	1.60
S	0.09	None (beta=0)	Mixed effects model	4	50	3.16
T	0.09	None (beta=0)	Mixed effects model	20	50	1.36
U	0.09	None (beta=0)	Mixed effects model	50	50	1.09
V	0.09	None (beta=0)	Mixed effects model	4	100	5.41
W	0.09	None (beta=0)	Mixed effects model	20	100	1.81
X	0.09	None (beta=0)	Mixed effects model	50	100	1.27
Y	0.01	None (beta=0)	Mixed effects model	4	50	1.24
Z	0.01	None (beta=0)	Mixed effects model	20	50	1.04
AA	0.01	None (beta=0)	Mixed effects model	50	50	1.01
BB	0.01	None (beta=0)	Mixed effects model	50	500	1.19

Scenario	Intraclass correlation (ICC)	Treatment effect	Model from which data were generated	Number of clusters (<i>m</i>)	Sample size per group (<i>n</i>)	Design effect 1+ICC(CS-1)
CC	0.001	None (beta=0)	Mixed effects model	4	50	1.024
DD	0.001	None (beta=0)	Mixed effects model	20	50	1.004
EE	0.001	None (beta=0)	Mixed effects model	50	50	1.001

Table 2

Description of 31 cluster definitions assumed when generating robust standard error estimates of treatment effect.

Cluster	Description	Number of clusters	Nested Within arm?	Cluster size
1	Each subjects is its own cluster	0 or N	Nested	Equal
2	4 equal-sized clusters, 2 nested in each treatment group	4	Nested	Equal
3	10 equal-sized clusters, 5 nested in each treatment group	10	Nested	Equal
4	30 equal-sized clusters, 15 nested in each treatment group	30	Nested	Equal
5	50 equal-sized clusters, 25 nested in each treatment group	50	Nested	Equal
6	4 equal-sized clusters across arms	4	Not nested	Equal
7	10 equal-sized clusters across arms	10	Not nested	Equal
8	30 equal-sized clusters across arms	30	Not nested	Equal
9	50 equal-sized clusters across arms	50	Not nested	Equal
10	4 unequal-sized clusters, 2 nested within each arm	4	Nested	Unequal
11	10 unequal-sized clusters, 2 nested within each arm	10	Nested	Unequal
12	30 unequal-sized clusters, 2 nested within each arm	30	Nested	Unequal
13	50 unequal-sized clusters, 2 nested within each arm	50	Nested	Unequal
14	4 clusters of unequal size across arms	4	Not nested	Unequal
15	10 clusters of unequal size across arms	10	Not nested	Unequal
16	30 clusters of unequal size across arms	30	Not nested	Unequal
17	50 clusters of unequal size across arms	50	Not nested	Unequal
18	4 clusters of unequal size across arms	4	NOT nested	Unequal and proportions vary by arm
19	10 clusters of unequal size across arms	10	NOT nested	Unequal and proportions vary by arm
20	30 clusters of unequal size across arms	30	NOT nested	Unequal and proportions vary by arm
21	50 clusters of unequal size across arms	50	NOT nested	Unequal and proportions vary by arm
22	2 clusters in 1 arm	$n+2$	Nested	Clustering in 1 arm only
23	4 clusters in 1 arm	$n+4$	Nested	Clustering in 1 arm only
24	10 clusters in 1 arm	$n+10$	Nested	Clustering in 1 arm only
25	30 clusters in 1 arm	$n+30$	Nested	Clustering in 1 arm only
26	True number of clusters, but item belongs to right cluster with probability 0.95	True number of clusters	Not nested	N/A
27	True number of clusters, but item belongs to right cluster with probability 0.90	True number of clusters	Not nested	N/A
28	True number of clusters, but item belongs to right cluster with probability 0.85	True number of clusters	Not nested	N/A

Cluster	Description	Number of clusters	Nested Within arm?	Cluster size
29	True number of clusters, but item belongs to right cluster with probability 0.80	True number of clusters	Not nested	N/A
30	True number of clusters, but item belongs to right cluster with probability 0.75	True number of clusters	Not nested	N/A
31	True number of clusters, but item belongs to right cluster with probability 0.70	True number of clusters	Not nested	N/A

Table 3

Results from fitting models to data generated with no clustering for $N=100$.

Cluster assumptions		Scenario A no treatment effect	Scenario D treatment effect
Cluster membership	Number of clusters	% reject null	% true effect contained in 95% CI
True Model	0	5.1	94.8
Cluster 1	0 or 100	5.4	94.7
Cluster 2	4	29.7	70.4
Cluster 3	10	12.6	88.4
Cluster 4	30	6.9	94
Cluster 5	50	6.3	94
Cluster 6	4	18.2	81.4
Cluster 7	10	9.3	90.2
Cluster 8	30	7.4	92.9
Cluster 9	50	6.6	94.3
Cluster 10	4	38.9	62.1
Cluster 11	10	16	84.3
Cluster 12	30	10.6	88.8
Cluster 13	50	8.4	92.1
Cluster 14	4	31.8	70.2
Cluster 15	10	15.3	86.1
Cluster 16	30	10.1	89.5
Cluster 17	50	8.1	91.3
Cluster 18	4	27.9	72.6
Cluster 19	10	12.7	84.3
Cluster 20	30	10.8	89.8
Cluster 21	50	8.1	91.2
Cluster 22	52	11.1	89.2
Cluster 23	54	8.6	91.1
Cluster 24	60	6.6	93.1

Cluster assumptions		Scenario A no treatment effect	Scenario D treatment effect
Cluster membership	Number of clusters	% reject null	% true effect contained in 95% CI
Cluster 25	80	5.9	93.4
		0.039	0.039

$\widehat{S}^a$ = average estimate of variance.

Table 4

Results from fitting models to data generated with high intra-cluster correlation (ICC=0.50) for different numbers of clusters and $N=100$.

Cluster assumptions		Scenario G ($m=4$)	Scenario H ($m=20$)	Scenario I ($m=50$)
Cluster membership	Number of clusters	% reject null $\widehat{S}_a$	% reject null $\widehat{S}$	% reject null $\widehat{S}$
True model	m	4.4	5.1	4.6
OLS model	0	63	29.3	17.3
OLS model with correct cluster specified	m	31.6	7.4	7
Cluster 1	0 or N	63.2	29.9	17.8
Cluster 2	4	76.3	54.6	45
Cluster 3	10	65.3	36.8	24.9
Cluster 4	30	63.2	31.9	21.3
Cluster 5	50	64	30.7	19.3
Cluster 6	4	70.8	44	34.9
Cluster 7	10	66.3	36	23
Cluster 8	30	63.2	32.2	19.6
Cluster 9	50	64.2	30.5	17.5
Cluster 10	4	81.6	62.1	51.7
Cluster 11	10	69.8	42.9	29.3
Cluster 12	30	67.3	38.5	25.5
Cluster 13	50	65.2	33.5	21.3
Cluster 14	4	76.5	55.8	43.5
Cluster 15	10	70.1	40.4	29.9
Cluster 16	30	67.1	36	24.6
Cluster 17	50	64.5	34.1	22.2
Cluster 18	4	76.5	50.9	41.8
Cluster 19	10	69.2	41	29.6
Cluster 20	30	66.2	36	25.3
Cluster 21	50	65.5	33.1	21.5
Cluster 22	$n+2$	67.6	38.3	27.6
Cluster 23	$n+4$	65.8	34.5	21.9

Cluster assumptions		Scenario G (<i>m</i> =4)	Scenario H (<i>m</i> =20)	Scenario I (<i>m</i> =50)
Cluster membership	Number of clusters	% reject null	$\overline{\overline{S^a}}$	% reject null
Cluster 24	<i>n</i> +10	63.6	0.056	19
Cluster 25	<i>n</i> +30	63.4	0.058	18.2
Cluster 26	<i>m</i>	33.9	0.448	7.6
Cluster 27	<i>m</i>	37.1	0.394	7.9
Cluster 28	<i>m</i>	39.7	0.341	9
Cluster 29	<i>m</i>	40.7	0.295	10.1
Cluster 30	<i>m</i>	43.5	0.249	11.3
Cluster 31	<i>m</i>	46.5	0.213	11.6

$\overline{\overline{S^a}}$ = average estimate of variance.

Table 5

Results from fitting models to data generated from a large number of clusters ($m=50$) with moderate to low ICCs (ICC=0.20, 0.09, 0.01, and 0.001) for $N=100$.

Clustering assumptions	Membership	m	Scenario O (ICC=0.20)		Scenario U (ICC=0.09)		Scenario AA (ICC=0.01)		Scenario EE (ICC=0.001)	
			% reject null	$\overline{S^a}$	% reject null	$\overline{S^a}$	% reject null	$\overline{S^a}$	% reject null	$\overline{S^a}$
True model		m	5.4	0.067	5.8	0.052	3.9	0.011	4.6	0.875
OLS model		0	10.5	0.05	7.3	0.044	4.8	0.01	5.7	0.818
OLS model with correct cluster specified		m	6.8	0.065	6.8	0.049	5.2	0.009	7.5	0.757
Cluster 1		0 or N	10.5	0.049	7.3	0.043	4.9	0.01	6	0.801
Cluster 2		4	38.7	0.023	31.8	0.022	31	0.005	30.2	0.401
Cluster 3		10	17.2	0.04	14.5	0.034	13	0.007	13.5	0.645
Cluster 4		30	13.7	0.045	9.5	0.04	8.1	0.009	8	0.751
Cluster 5		50	11.9	0.046	8.5	0.041	6.1	0.009	7.5	0.76
Cluster 6		4	24	0.036	19.6	0.033	18.1	0.007	20.9	0.58
Cluster 7		10	14.9	0.044	11.5	0.039	9.8	0.009	10.9	0.74
Cluster 8		30	12.7	0.046	8.2	0.042	6.8	0.009	7.6	0.778
Cluster 9		50	11.6	0.047	8.3	0.042	5.8	0.009	7.3	0.784
Cluster 10		4	45.6	0.015	39.4	0.014	39.1	0.003	37.6	0.256
Cluster 11		10	21.7	0.032	16.4	0.03	14.7	0.007	16.8	0.546
Cluster 12		30	16.4	0.038	11.7	0.034	10.1	0.008	11.3	0.631
Cluster 13		50	13.6	0.043	9.7	0.038	8.8	0.008	10.1	0.708
Cluster 14		4	34.9	0.024	28.7	0.021	30.5	0.004	29	0.386
Cluster 15		10	20.1	0.035	17.7	0.03	16.3	0.007	13.9	0.59
Cluster 16		30	16.3	0.04	13.1	0.034	10.6	0.008	11.5	0.646
Cluster 17		50	14.2	0.044	9.9	0.038	9	0.008	10	0.722
Cluster 18		4	35.1	0.023	29.5	0.02	29.3	0.005	28	0.38
Cluster 19		10	20.3	0.036	16	0.032	15.3	0.007	15	0.597
Cluster 20		30	16.4	0.039	13.1	0.034	9.7	0.008	10.7	0.646
Cluster 21		50	13.6	0.043	10.5	0.038	7.7	0.009	9.3	0.723
Cluster 22		$n-2$	18.3	0.036	14.2	0.033	11	0.007	12.9	0.595

Clustering assumptions		Scenario O (ICC=0.20)		Scenario U (ICC=0.09)		Scenario AA (ICC=0.01)		Scenario EE (ICC=0.001)	
Membership	m	% reject null	$\overline{S^a}$	% reject null	$\overline{S^a}$	% reject null	$\overline{S}$	% reject null	$\overline{S}$
Cluster 23	$n+4$	15.6	0.042	11	0.037	8.9	0.008	9.3	0.704
Cluster 24	$n+10$	12.1	0.046	9.1	0.04	6.8	0.009	7.9	0.76
Cluster 25	$n+30$	10.5	0.048	8.2	0.042	5.4	0.009	7.5	0.791
Cluster 26	m	7.3	0.063	6.8	0.048	5.6	0.009	7.6	0.756
Cluster 27	m	7.2	0.061	7.2	0.047	5.2	0.009	6.9	0.764
Cluster 28	m	7.8	0.06	7.6	0.047	5.5	0.009	7.3	0.769
Cluster 29	m	8.6	0.058	7.8	0.046	5.4	0.009	7.1	0.768
Cluster 30	m	9.1	0.057	7.5	0.046	6.4	0.009	7.8	0.776
Cluster 31	m	8.9	0.056	7.4	0.046	6.2	0.009	7.7	0.77

$\overline{S^a}$ = average estimate of variance.